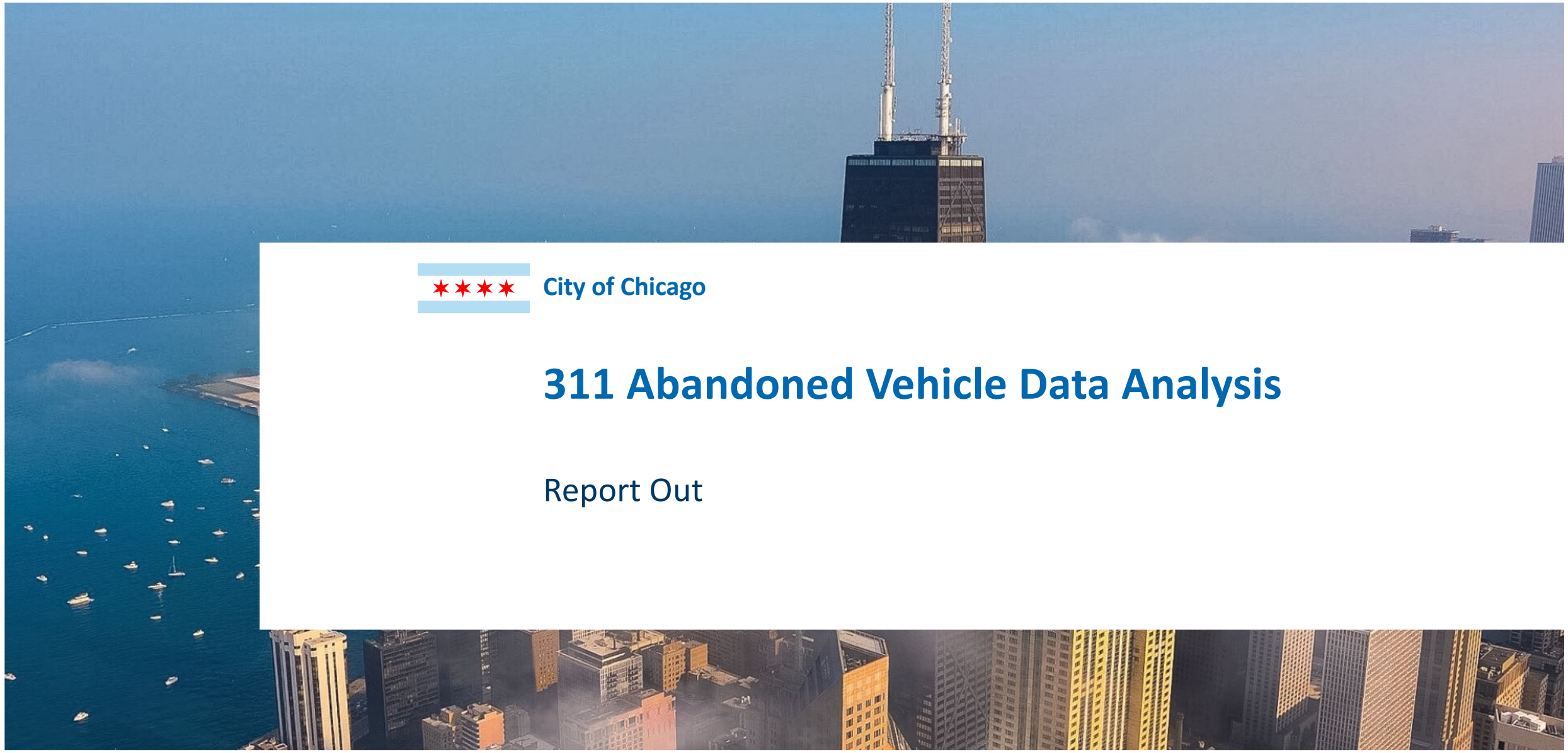


City of Chicago

311 Abandoned Vehicle Data Analysis

Report Out

Contributor: Aurora Peng

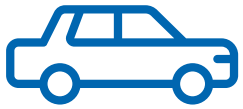
June 2022

Agenda

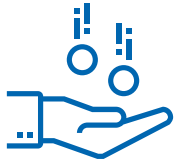
1. Background & High-Level Approach
2. Exploratory Data Analysis
3. Data Preprocessing
4. Model Evaluation and Selection
5. Project Wrap-up

Problem Statement

The City of Chicago continues having issues with abandoned cars and lacks analytical insights to determine paths to reduce response times, allocate service distributions and improve financial efficiencies



Abandoned cars are a continuous issues in urban areas across the nation, especially in Chicago due to extreme population density



Having **abandoned cars leads to social issues** which **costs the city and taxpayers money** due to increase in hazardous waste, littering, general blight, etc.

Chicago lacks analytical insights on the abandoned car data that would allow the City to improve its services for residents

I aim to serve the City of Chicago to perform an **end-to-end data pipeline and business analytics** to provide a new EDA feature that would allow the citizen to better **expect the time to complete the service and to better relocate the service**

High-Level Approach

I employed a systematic 5-step approach to provide the Citizen of Chicago with a set of data-driven recommendations (based on foundational datasets the City made available)

1. Data Gathering

- 311 Abandoned Car Reports

2. Exploratory Data Analysis

- Remove empty records (Nulls)
- Limit outlier values
- Identify duplicate entries

3. Data Preprocessing

- Use SimpleImputer etc to impute NA
- Use OneHotEncoder etc to impute the categorical features

4. Data Modeling

- Linear Regression Model
- Lasso Regression Model
- Ridge Regression Model

5. Model Evaluation

- See the potential reason for the failure

1. What is the dataset looks like?

	Vehicle Make/Model	Vehicle Color	Current Activity	Most Recent Action	ZIP Code	Ward	Police District	Community Area	Latitude	Longitude	creation_year	creation_month	creation_day	Completion_I
0	NaN	NaN	NaN	NaN	60619.0	6.0	3.0	69.0	41.761761	-87.620146	2013	1	8	
1	NaN	NaN	NaN	NaN	60601.0	42.0	1.0	32.0	41.884262	-87.617296	2011	1	13	
2	NaN	RED	NaN	NaN	NaN	37.0	25.0	25.0	41.909527	-87.745257	2011	1	14	
3	NaN	NaN	NaN	NaN	60601.0	42.0	1.0	32.0	41.887331	-87.617133	2012	1	20	
4	NaN	NaN	NaN	NaN	60644.0	28.0	15.0	25.0	41.877796	-87.745662	2011	1	24	

Observations

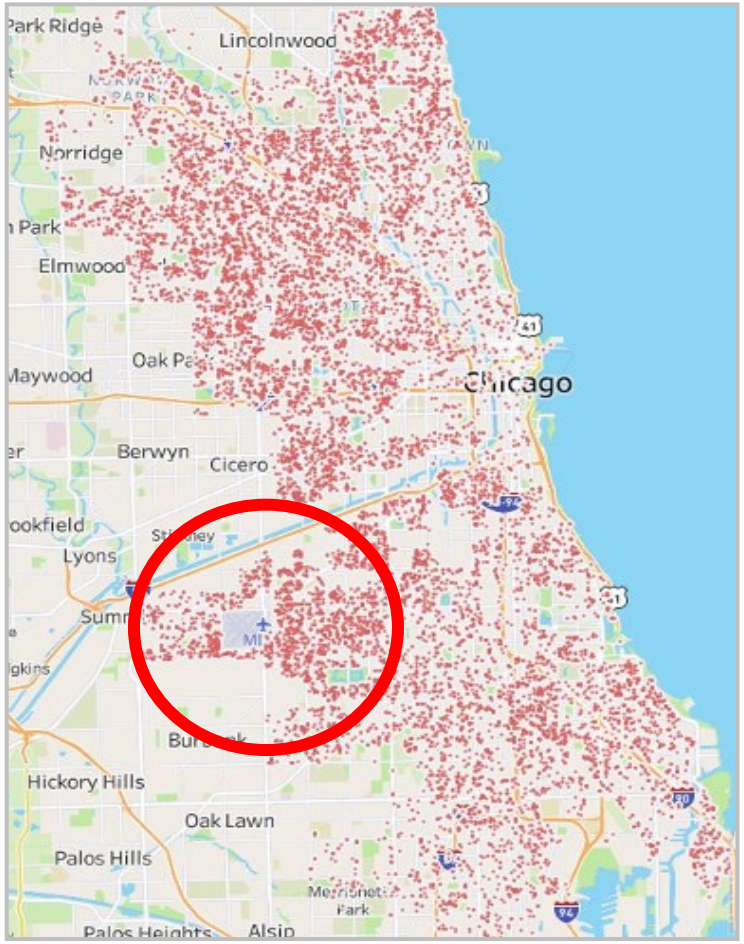
- Dataset: 231608 entries
- Train Test: using train test split to get the train and test dataset

Features

- 13 features
- 4 categorical features :'Vehicle Make/Model', 'Vehicle Color', 'Current Activity', 'Most Recent Action’
- 9umerical features :'ZIP Code’, 'Ward’, 'Police District’, ’ Community Area','Latitude','Longitude','creation_year','creation_month','creation_day'

2. What is the best neighborhood to abandon my car in Chicago?

Occurrences of First Abandoned Car Reports (2020)



Notes

Observations

- In 2020, there was a high concentration of abandoned car reports around Midway airport

Takeaways

- Areas around Midway airport are suitable to stash or abandon your car(s)

Data Preprocessing Approach – NA/Categorical Data

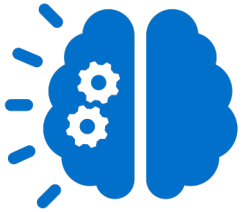
We are preprocessing the data to fit into the future data modeling and the prediction

NAs

For the numerical NAs:

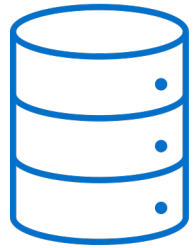
- *Filling the numerical NAs with the mean values*
- *Filling the categorical NAs with the most common classes*
- *Using SimpleImputer and lamda functions*

Categorical Features



1. **One hot encoder:**
   - for the multiclass features we are using one hot encoders
   - Multiclass features: 'Vehicle Make/Model', 'Vehicle Color', 'Most Recent Action'
2. **Ordinal Encoder:**
   - For the ordinal class features, we are using ordinal encoder
   - Ordinal class features: 'Current Activity'

Scaling



MaxAbs Scaler

- *We are using Max Abs scaler because we want to minimize the impact from the outliers and the noises*
- *By using Max Abs scaler we can derive all the features into a specific training range*

Data Modeling

For the regression problem, we are fitting the data into the simple regression, linear regression, regularized regression models.

1. **Model Purpose**: Add an ETA feature to predict the estimated complete service date provided to the citizens

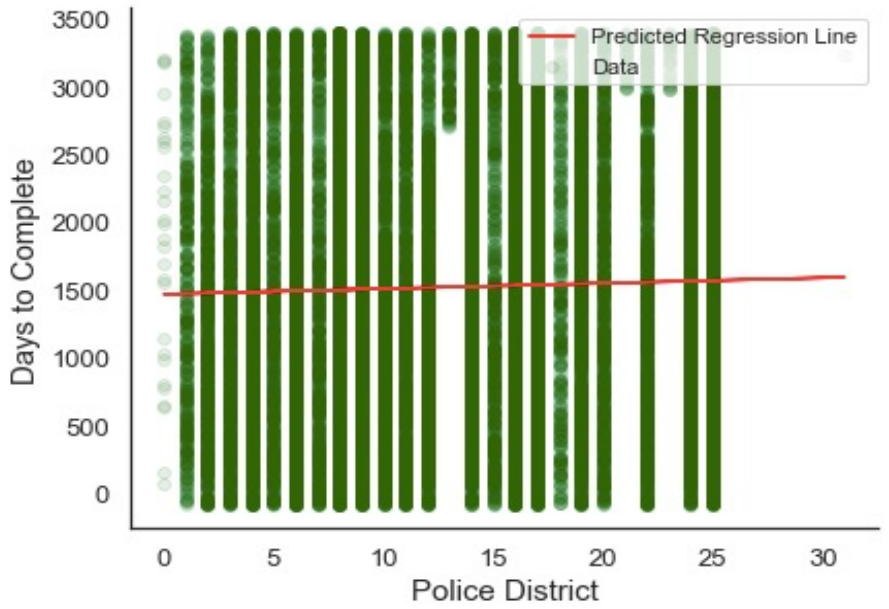
2. **Data Cleaning**: Use Record Linkage package to delete the duplicate

3. **Model Selection**: Simple/Multiple Linear Regression, Ridge/Lasso Linear Regression, Polynomial Regression

Note:

Obviously, this is far from precise for a simple linear regression for the correct value.

Let's figure out some regression with the whole features!



Data Modeling

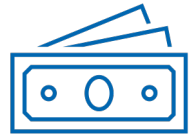
For the regression problem, we are fitting the data into the simple regression, linear regression, regularized regression models.

	Model	Details	Root Mean Squared Error (RMSE)	R-squared (training)	Adjusted R-squared (training)	R-squared (test)	Adjusted R-squared (test)	5-Fold Cross Validation
1	Multiple Regression-4	all features	987.905	0.001	-0.0	0.001	-0.004	1.000
2	Ridge Regression	alpha=1, all features	988.344	0.001	-0.0	0.001	-0.004	1.000
3	Ridge Regression	alpha=100, all features	988.344	0.001	-0.0	0.001	-0.004	1.000
4	Ridge Regression	alpha=1000, all features	988.344	0.001	-0.0	0.001	-0.004	1.000
5	Lasso Regression	alpha=1, all features	988.344	0.001	-0.0	0.001	-0.004	1.000
6	Lasso Regression	alpha=100, all features	988.344	0.001	-0.001	0.000	-0.004	0.887
7	Lasso Regression	alpha=1000, all features	988.344	0.000	-0.001	-0.000	-0.005	-7.620
0	Simple Linear Regression	-	988.344	0.001	-	0.001	-	-27.270

Potential Reasons of Failed Prediction

The kernel died after the polynomial regression. Poly can somehow give me a better solution since the complexity of the model do grows up which leads to the higher variance.

The outlier processing is raw. The dataset is really raw and I haven't processed the outlier outside of distribution of 95%. I believe by process this outliers the model will be more accurate and concise.



The lack of grid search. Due to the lack of the dataset, I haven't done the grid search which can definitely miss some of the models in the best practice.

Last but not least,

- If you want to **stash your car**, do it **near Midway Airport**!
- If you are considering **buying a new car**, **AVOID a green Dodge**